Review of: "Mathematical and Linguistic Characterization of Orhan Pamuk’s Nobel Works"

Martina Benešová[1]

1 Palacky University

Potential competing interests: No potential competing interests to declare.

The paper presents an exploration into the potential of quantitative-linguistic methods, particularly those probing the fractal properties of texts, through an analysis of works by Orhan Pamuk. The authors meticulously outline four distinct approaches in their methodology, providing detailed descriptions accompanied by visual representations in tables and graphs for enhanced comprehension (these visual aids serve above all to facilitate understanding, not only to *make the data mathematically meaningful*, as noted in the paper). The choice of basic units for analysis—letters (i.e., graphemes, not *sounds* as mentioned in the paper) and words (i.e., graphical words or tokens, later expanded to lemmas)—is reasonable given the focus on written text. However, it is suggested that the authors clearly define these units before commencing the analysis, a step that would further strengthen the precision of their study.

The authors provide a thorough background overview in the Introduction, delving into various disciplines such as referential literature, fractal qualities of music, Eftekhari’s work, and aphasias, setting a solid foundation for their study. They draw a parallel between the concept of fractals in music, where notes serve as basic units, and extend this analogy to texts, where letters and words are considered equivalent to notes. This transition from music to textual analysis is an intriguing approach that offers a fresh perspective on the study of fractality in linguistic structures.

At the beginning of the Methods and Results section, the authors express that *in general,* [their] *aim is to show that a mathematical analysis of literature, which can provide* [them] *with valuable quantitative values, can be discussed and speculated in the semantic world*. It might be beneficial to clearly articulate a plain hypothesis to guide the study’s direction. The authors’ aspiration is to test employing quantitative methods for textual analysis, as well as to focus on the works of a specific author, which is commendable yet ambitious. However, such an approach can be intricate when dealing with numerous boundary conditions. Additionally, the authors allude to the interpretation within the semantic world and the identification of emotions and thoughts through text analytics, which may be challenging to achieve at this stage of research; the reader does not achieve such interpretation, but hopes to gain it in the future.

The methodology for handling text is well-explained; however, readers might benefit from additional details regarding text preprocessing. Such analyses are notably sensitive to initial conditions. For instance, readers might appreciate insights into how numbers are handled, whether stop words are included or excluded, and other preprocessing steps. Furthermore, it is advisable to utilize standard terminology in the text processing field, such as “frequency” instead of “*number of times*,” and terms like “lemma” and “lemmatization,” “tokenization,” etc. These standard terms will enhance the clarity and accessibility of the methodology section.

In the first approach, where the authors consider letters as the basic units, they order them in two ways: by their frequencies and alphabetically. The authors themselves acknowledge that alphabetical ordering is artificial and may not accurately reflect the quality of the text. Therefore, it raises questions about its value in comparison with frequency ordering.

The paper briefly mentions the concept of self-similarity, yet its application in the methodology remains unclear. For instance, Hřebíček (1995), referenced by the authors, discusses self-similarity in a heuristic and somewhat mathematically controversial yet explicit manner, based on the similarity between the whole and its parts. However, it is not explicitly outlined how this concept is detected or utilized within the methodology presented in this paper. The introduction of Zipf's law is beneficial, but its explicit relation to fractality is not clearly articulated. While the authors note the similarity between the formulas of Zipf's law and the fractal dimension, this alone does not establish a text as a fractal whose dimension can be measured. The comparison of fractal dimension D_Z to well-known dimensions like the Koch curve and the coastline of Britain is intriguing, but it leaves open the question of whether this similarity is more than just a coincidence.

Moreover, the discussion raises questions about the independence of text dimensions from the text length (as mentioned in the paper), or whether the length contributes to the precision or statistical reliability of the outputs, thereby indicating a better goodness-of-fit. The statement about Zipf's dimensions of other novels being "almost *linear*" is puzzling, as dimensions themselves cannot be linear. It would be beneficial for the authors to delve deeper into these concepts and perhaps explore analyses of other texts, authorial styles, and languages to provide a more comprehensive understanding of the relationships between dimensions and textual features.

The paper also delves into the analysis based on words, later specifically highlighting a subset of prepositions, although it is confusingly noted that conjunctions are also utilized in Figures 9 and 10. Providing more insight into why prepositions were chosen over other word types would offer a deeper understanding of the methodology and its implications. Analyses focusing on other word groups, or e.g., hapax legomena, would be a beneficial contribution to further research.

In addition to comments on the content, there are also observations regarding the formal structure of the paper. The reader may notice occasional misspellings or errors, such as "*fractals have been interested in music*", a repeating sentence ("*For each novel, a frequency list of words has been prepared*"), and inconsistent use of hyphens and spaces. There is also an example of "*much different than*" that could be improved for clarity. These issues detract slightly from the overall readability of the paper and should be addressed for a more polished presentation.

One of the main challenges for readers of this paper is the presentation of visualizations and tables. It is highly recommended to include concise yet comprehensive descriptions alongside these visuals to ensure readers can fully grasp their meaning without needing to refer back to the text. For example, in the case of graphs, it should be clear which variables are being depicted and under what specific conditions, as well as the significance of observation points and curves. Therefore, it is essential to label axes clearly to prevent confusion. Without such guidance, readers may struggle to interpret the information presented, as seen in Figure 4, where the meaning of data points is not immediately apparent. Additionally, it is important to note that a dataset with only four data points may not be statistically robust, so it would be

valuable to consider or explain its significance in the context of the research.

To enhance the reader-friendliness of the paper, it is advisable to organize the figures in a more coherent manner to avoid readers needing to flip back and forth between pages. For example, Figures 8 to 11 contain 12 graphs clustered together in a continuous sequence with the explanatory text placed separately. Additionally, these figures initially present frequency visualizations followed by their log-log versions (with semi-log plotting mentioned in the Discussion). This raises questions: Why are both versions included? Is this not essentially duplicating the visualization from different perspectives? Taking the logarithm of both frequency and rank serves to transform the data into a linear relationship, smoothing out the distribution and facilitating visualization, identification, and assessment of power law behavior.

It is important to mention the conventional presentation of Zipf's law, which typically displays decreasing frequencies on the y-axis and increasing rank on the x-axis. This format is immediately recognizable to researchers familiar with the concept. Therefore, the alternative presentation method used in this paper warrants a brief remark.

Figure 6 and the accompanying text present Zipf's orders for all four novels under consideration, noting their striking similarity. However, it is advisable to first investigate whether this similarity is a characteristic of the author's work, the Turkish language, or Turkic languages in general.

Table 4, which is mentioned in the paper to *show the types and number of fiction texts compiled in the construction of the CCTF*, actually presents *Zipf's dimensions for letters of different novels by Pamuk,* as is written in its caption. This raises the question of whether any figure is missing from the paper.

In the Discussion, the authors present not only mathematical interpretations of the obtained outputs but also outputs of further tests and results, yet no supportive graphs, tables, coefficients of determination, or correlation coefficients are supplied. This section introduces an important possibility of comparative frequency analysis, which, nevertheless, would benefit from more extensive treatment with additional quantitative and statistical methods, rather than a brief assessment (only the first two ranks and one more significant word are treated, and it is unclear if the frequencies relate to types or tokens). It may be recommended to consider involving other methods or focuses of text processing, such as N-grams or hapax legomena, or to perform a Part-of-Speech (POS) or keyword analysis for more comprehensive insights.

The authors promise in their paper to pursue follow-up research, and the field of text fractality is undoubtedly worth further exploration. It would be intriguing, for example, to delve into why the dimensions of letters fall within the interval of <1;2> and those of words within <0;1> (or if this trend is generally valid across languages and styles). Exploring the qualities of different languages and styles using the outlined methodology, or examining different language levels, could also yield valuable insights.

It is worth noting that twice in the paper, the authors mention "*fractal and linguistic languages*," a concept that would benefit from a clear definition and thorough explanation. This clarification would help readers understand the authors' intended meaning and its relevance to their study.

Last but not least, the paper focuses on a linguistic research problem, suggesting that a linguistic interpretation could

greatly enhance its potential.